RNA Sequencing

DOI: 10.1002/anie.200906336

Reading the Code of Single RNA Molecules

*Sabine Müller**

high-throughput screening · RNA · sequence determination · single-molecule studies · transcriptome

The advent of DNA sequencing in the late 1970s[1] has fundamentally altered life sciences. The chain termination method by Sanger,[1c] also referred to as dideoxy method, has dominated decades of sequence-driven research, with the completion of the human genome being a major milestone.[2] Nowadays, the Sanger method has been partially replaced by next-generation sequencing technologies, such as 454/Roche,[3] Illumina/Solexa,[4] ABI/SOLiD,[5] Helicos,[6] and Pacific Biosciences sequencing.[7] All of these new generation technologies follow the principle of sequencing by synthesis, and offer dramatic increase in cost-effective throughput, although at the expense of read length. Therefore, next-generation sequencing technologies, and in particular 454 sequencing, have been used predominantly for cDNA sequencing in transcriptome analysis, rather than for sequencing of genomic DNA.

The in vivo amplification required in Sanger sequencing is circumvented by cloning-free in vitro amplification of spatially separated DNA molecules, known as emulsion PCR[8] (454/Roche and ABI/SOLiD) or by solid-phase bridge amplification of single-molecule DNA templates (Illumina/Solexa). As a result, colonies of DNA templates are produced that are immobilized on a surface, sequenced, and analyzed. The protocols developed by Helicos[6] and Pacific Biosciences[7] are based on single-molecule technologies, such that no amplification is required. The Helicos protocol[6] uses primer-template duplexes that are immobilized on glass cover slips, and single-molecule sequencing is achieved by addition of labeled nucleotides in a step-wise fashion, followed by imaging of the colored positions on the slip after each round of nucleotide addition.

This DNA sequencing protocol now has been adapted to direct RNA sequencing (DRS).[9] This is a major achievement, because DRS has the potential to directly sequence femtomole amounts of total RNA from any given cell population without prior copying to cDNA.

This achievement greatly enhances research into gene expression profiling, genome annotation, and rearrangement detection to non-coding RNA discovery and quantification.

[*] Prof. Dr. S. Müller
Ernst-Moritz-Arndt Universität Greifswald, Institut für Biochemie
Felix-Hausdorff-Strasse 4, 17487 Greifswald (Germany)
Fax: (+49) 3834-864-471
E-mail: smueller@uni-greifswald.de
Homepage: http://www.chemie.uni-greifswald.de/~bioorganik

Like the Helicos DNA sequencing procedure,[6] DRS is a single-molecule method. In the first step, *E. coli* poly(A) polymerase I (PAPI) is used to generate an A tail of about 150 nucleotides at the 3′-end of RNA molecules to be analyzed, whereby RNAs that naturally contain poly(A) tails, such as mRNAs, may be excluded. Control of the tail length and blocking of the 3′-end to prevent downward addition of nucleotides in the sequencing reaction is achieved by addition of ddATP to the tailing reaction ten minutes after reaction start. Templates are hybridized to poly(dT)-coated glass cover slips and analyzed by stepwise addition of fluorescently labeled and 3′-blocked nucleotides, called virtual terminator nucleotides (VT nucleotides). To define the sequencing start point, RNA templates are filled in with dTTP and locked in position by addition of VT-A, VT-C, or VT-G (Figure 1A). Unincorporated VT nucleotides are removed by washing, and the localization of the label on the chip is then imaged. The fluorescence dye and inhibitor group is then cleaved off, rendering the 3′-OH group free for addition of the next nucleotide (Figure 1B). The four nucleotides (VT-A, VT-C, VT-G, and VT-T) are then added in alternating order, followed by washing, imaging, and cleavage (Figure 1C–F). After iterative rounds of these steps, images are aligned and used to generate the sequence of each individual RNA molecule.

The key elements of this procedure are the single-molecule technique, a polymerase that accepts modified fluorescently labeled nucleotides as substrates, and the design of the 3′-blocked labeled nucleotide. Single-molecule sequencing was proposed as early as 1989,[10] but its feasibility has been demonstrated only recently.[11] Apart from the fact that only minute amounts of material are required, a big advantage of single-molecule sequencing is that each molecule is monitored individually. Thus, nucleotide incorporation does not need to be driven to completion, which in turn reduces misincorporations, and thus the error rate. Owing to their slow reaction kinetics, non-complementary nucleotides cannot compete with the time required to incorporate 80 to 90% of the correct base.[6,9]

To incorporate the labeled VT nucleotides, a modified polymerase is required. Ozsolak et al.[9] screened known reverse transcriptases, and also several DNA-dependent DNA polymerases that had been previously shown to have an engineered reverse-transcriptase activity.[12] Obviously, they have identified a polymerase that efficiently incorporates the VT nucleotides, thereby differentiating between the

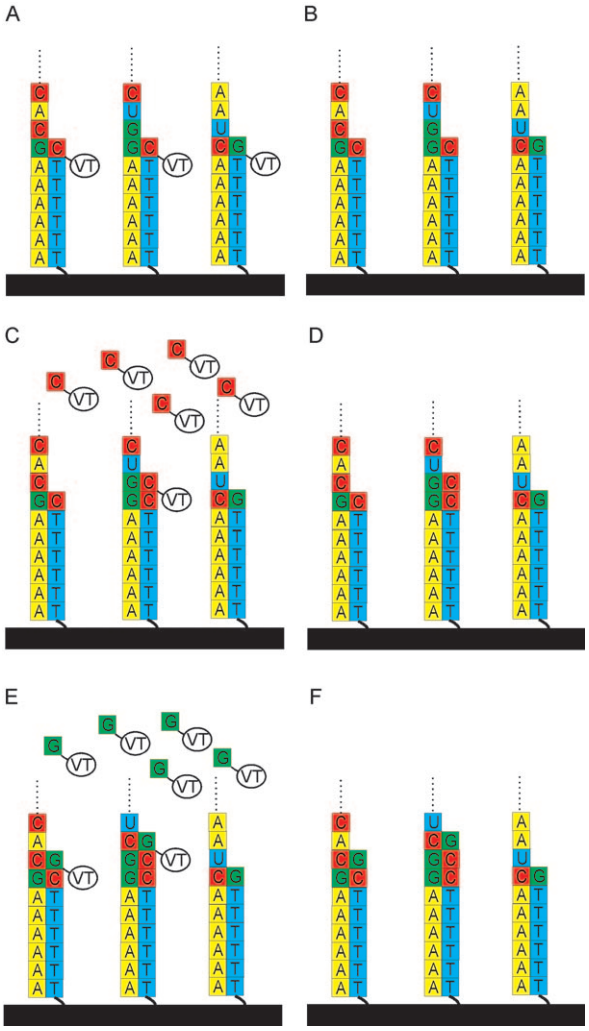
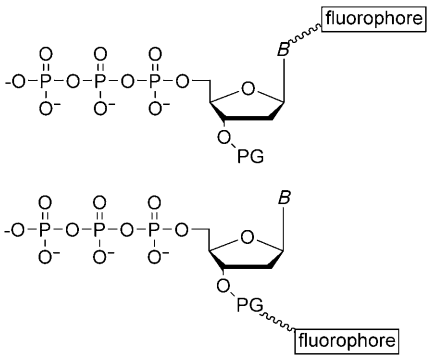


Figure 1. RNA sequencing. A) Polyadenylated templates are hybridized to a poly(dT)-coated surface, filled in with dTTP, and locked in position with VT nucleotides. B,D,F) The fluorescence dye and the 3′-*O* blocking group are cleaved off, generating a free 3′-OH group. C,E) VT nucleotides (for example, VT-C or VT-G) are added, followed by washing and imaging the position of the label on the chip.

correct and the impaired nucleotide sufficiently well. Disappointingly, the paper does not give any further information on the nature of the polymerase; the same applies to the VT nucleotides. It is left to the fantasy of the reader as to what the chemical nature of these nucleotides is, and how the dye and the inhibitor group are cleaved off.

One possibility is using fluorescently labeled 3′-*O*-blocked nucleotides, with the fluorescent label attached to the heterocyclic base (Scheme 1, top), and cleavage of the fluorescent label and of the 3′-*O*-blocking group under identical conditions.[13] This strategy is reasonable, as polymerases have been shown to accept nucleotides with bulky attachments.[14] Presumably, the VT-nucleotides used by Ozsolak et al.[9] are also decorated with a dye attached to the heterocyclic base and a separate 3′-*O*-blocking group, and both are cleaved under identical conditions. The dye is probably identical in all four VT nucleotides, because after each nucleotide incorporation step, the localization of the



Scheme 1. Possible structure of VT nucleotides. The fluorescent label may be attached to the heterocyclic base *B*, requiring an additional 3′-O-blocking group (top), or the fluorophore is directly linked to the 3′-O-blocking group (PG; bottom).

label on the chip is imaged rather than measuring the color of the released dye.

Alternatively, the fluorescent label might be attached directly to the 3′-*O*-blocking group (Scheme 1, bottom), such that efficient 3′-deblocking is monitored by the leakage of fluorescence. It is questionable whether natural polymerase would accept a nucleotide with a rather bulky group at the 3′-position as substrate; however, the technologies available nowadays for protein mutagenesis and evolution offer the chance to develop such polymerases.

With the new DRS technology at hand, Ozsolak et al.[9] first sequenced chemically synthesized 40-mer oligonucleotides as model systems to develop and optimize DRS chemistry, followed by sequencing *Saccharomyces cerevisiae* poly(A)$^+$ RNA, and alignment of the sequence reads to the yeast genome using bioinformatic tools. The average aligned read length was 28.7 nucleotides in 41261 reads, with 120 sequencing cycles over three days. This result puts DRS, though still in its infancy, in amongst high-throughput technologies, with the potential for exciting applications. After further advances and refinement, the method could be used to make snapshots of the transcriptome of any given population of cells for example, or even of individual cells. Unfortunately, the paper by Ozsolak et al.[9] does not provide information on the chemistry behind this technology and other scientific details. This absence may be understandable from the commercial point of view, yet it is disappointing to the reader, who, studying a publication in a scientific journal, expects to find all the information needed to fully understand the work.

Received: November 10, 2009
Published online: January 13, 2010

[1] a) F. Sanger, A. R. Coulson, *J. Mol. Biol.* **1975**, *94*, 441–448; b) A. M. Maxam, W. Gilbert, *Proc. Natl. Acad. Sci. USA* **1977**, *74*, 560–564; c) F. Sanger, S. Nicklen, A. R. Coulson, *Proc. Natl. Acad. Sci. USA* **1977**, *74*, 5463–5467.
[2] a) E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al., *Nature* **2001**, *409*, 860–921; b) International Human

Genome Sequencing Consortium, *Nature* **2004**, *431*, 931 – 945; c) J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, et al., *Science* **2001**, *291*, 1304 – 1351.

[3] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, et al., *Nature* **2005**, *437*, 376 – 380.

[4] a) S. T. Bennett, *Pharmacogenomics* **2004**, *5*, 433 – 438; b) S. T. Bennett, C. Barnes, A. Cox, L. Davies, C. Brown, *Pharmacogenomics* **2005**, *6*, 373 – 382; c) D. R. Bentley, *Curr. Opin. Genet. Dev.* **2006**, *16*, 545 – 552.

[5] J. Shendure, G. J. Porreca, N. B. Reppas, X. Lin, J. P. McCutcheon, A. M. Rosenbaum, M. D. Wang, K. Zhang, R. D. Mitra, G. M. Church, *Science* **2005**, *309*, 1728 – 1732.

[6] T. D. Harris, P. R. Buzby, H. Babcock, E. Beer, J. Bowers, I. Braslavsky, M. Causey, J. Colonell, J. DiMeo, J. W. Efcavitch, et al., *Science* **2008**, *320*, 106 – 109.

[7] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, et al., *Science* **2009**, *323*, 133 – 138.

[8] D. S. Tawfik, A. D. Griffiths, *Nat. Biotechnol.* **1998**, *16*, 652 – 656.

[9] F. Ozsolak, A. R. Platt, D. R. Jones, J. G. Reiffenberger, L. E. Sass, P. McInerney, J. F. Thompson, J. Bowers, M. Jarosz, P. M. Milos, *Nature* **2009**, *461*, 814 – 818.

[10] J. H. Jett, R. A. Keller, J. C. Martin, B. L. Marrone, R. K. Moyzis, R. L. Ratliff, N. K. Seitzinger, E. B. Shera, C. C. Stewart, *J. Biomol. Struct. Dyn.* **1989**, *7*, 301 – 309.

[11] a) I. Braslavsky, B. Herbert, E. Kartalov, S. R. Quake, *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 3960 – 3964; b) W. J. Greenleaf, S. M. Block, *Science* **2006**, *313*, 801.

[12] a) S. Vichier-Guerre, S. Ferris, N. Auberger, K. Mahiddine, J.-L. Jestin, *Angew. Chem.* **2006**, *118*, 6279 – 6283; *Angew. Chem. Int. Ed.* **2006**, *45*, 6133 – 6137; b) K. B. M. Sauter, A. Marx, *Angew. Chem.* **2006**, *118*, 7795 – 7797; *Angew. Chem. Int. Ed.* **2006**, *45*, 7633 – 7635.

[13] D. C. Knapp, A. Keller, J. D'Onofrio, A. Lubys, S. Serva, A. Kurg, M. Remm, M. Kwiatkowski, J. W. Engels, *Nucleic Acids Symp. Ser.* **2008**, *52*, 345 – 346.

[14] a) S. Jäger, G. Rasched, H. Kornreich-Leshem, M. Engeser, O. Thum, M. Famulok, *J. Am. Chem. Soc.* **2005**, *127*, 15071 – 15082; b) S. H. Weisbrod, A. Marx, *Chem. Commun.* **2008**, 5675 – 5685.

